First, let’s break a person down into inputs, compute, and outputs.

Inputs:

- 2 decent 126MP cameras. “The human retina contains about 120 million rod cells, and 6 million cone cells” [1]
- One of those fancy two ear ASMR mics.
- Other sensors we don’t care about.

Outputs:

- A keyboard and mouse
- Other outputs we don’t care about.

Both inputs and outputs are cheap and simple with today’s tech, so we won’t talk about them further. It’s really all about the compute

---

There’s a lot of talk of the brain being complex. And while it’s true that it has complex behaviors, the computational substrate is quite easy to understand. A neuron is an accumulator, and a synapse is a multiplier. Multiply-accumulate. Sound familiar?

Each synapse is both a FLOP and a weight. In ANN’s, this isn’t always true with weight sharing, but since the compute is the memory in the brain, there’s no advantage for explicit weight sharing, and other weight syncing procedures are fine.

There’s varying estimates for the number of neurons and synapses in the brain, [2] claims 86 billion and 150 trillion. Other sites claim other things. [3] claims a child has 1000 trillion neurons, and the fanout is 7000. Of note, the 1000-10000x ratios of neurons to synapses give us an idea of the size of the weight matrices.

We’ll use the estimates of 100 billion neurons and 100 trillion synapses for this post. That’s 100 teraweights. GPT-3 has 175 gigaweights, so this brain is ~1000x bigger.

The max firing rate seems to be 200hz [4]. I really want an estimate of “neuron lag” here, but let’s upper bound it at 5ms. If reaction time is 200ms and “recognizing and responding to a visual stimulus” [5] takes 500ms, the processing is going through 20-100 serial layers.

Multiplying, this yields 20 PFLOPS of compute, and 200 TB of float16 weights.

nVidia’s new A100 claims 312 TFLOPS of compute, but unless you are weight sharing, you’ll run into the RAM bandwidth limit of 1.6 TB/s long before that. With float16, let’s round that to 1 TFLOP per GPU.

Remember also, that the brain is always learning, so it needs to be doing forward and backward passes. I’m not exactly sure why they are different, but [6] and [7] point to the backward pass taking 2x more compute than the forward pass.

We are up to 60 PFLOPS on **40,000 GPUs** (need to read and write the weight only). Since deep learning is usually done in minibatches, the GPU comes with 40 GB of RAM each, giving a total of 1.6 PB of RAM, a 8x overprovision.

With RAM bandwidth being the limiting factor, we get:

- 1x the RAM bandwidth
- 8x the storage (minibatch size)
- 312x the compute (weight sharing ratio)

nVidia is selling the GPUs for $12,500 each [8], so if you ordered from them, it would be $500 million. Though when you look at the chip cost, 54 billion transistors on 7 nm should only cost about $150. I just approved the purchase of 32GB RAM sticks[10] at $102 per, so RAM is about $3 per GB, meaning $120 for the card. $30 for the rest + $200 overhead for the machine they go in. $500 each x 40,000 is $20 million dollars.

Assuming they draw 400W each, that’s 16 MW of power. At $0.10 per kWh, it’s $1600 an hour to run.

$20 million + $1600 an hour

Citations

1. https://en.wikipedia.org/wiki/Photoreceptor_cell
2. https://en.wikipedia.org/wiki/List_of_animals_by_number_of_neurons
3. https://en.wikipedia.org/wiki/Neuron#Connectivity